PATTERN SEARCHING METHODS AND APPARATUSES

BACKGROUND OF THE INVENTION

1. Field of the Invention

The invention relates generally to a system and method for extracting relevant information from raw text data. More particularly, the invention concerns itself with a system and method for identifying patterns in text using structures defining types of patterns. In this context a "pattern" is to be understood as a part of a written text of arbitrary length. Thus, a pattern may be any series of alphanumeric characters within a text. Particular examples of patterns that might be identified in a text, such as a word-processor document or an email-text, are dates, events, numbers such as telephone numbers, addresses or names.

2. Description of the Background Art

Technologies for searching interesting patterns in a text presented by a computer to a user (in the following "computer text") are well-known. U.S. Pat. No. 5,864,789 is one example of a document describing such a technology.

A system that searches patterns in a computer text and provides to the user some actions based on the kind of identified patterns is described in two variants The first variant is an application termed "AppleDataDetectors" and the second variant an application termed "LiveDoc".

Both variants use the same method to find patterns in an unstructured text. The engine performing the pattern search refers to a library containing a collection of structures, each structure defining a pattern that is to be recognized. FIG. **1** gives an example of seven different structures (#**1** to #**7**), which may be contained in such a structure library. Each of the seven structures shown in FIG. **1** defines a pattern worth recognizing in a computer text. The definition of a pattern is a sequence of so-called definition items. Each definition item specifies an element of the text pattern that the structure recognizes. A definition item may be a specific string or a structure defining another pattern using definition items in the form of strings or structures. For example, structure #**1** gives the definition of what is to be identified as a US state code, the definition following the ":=" sign. According to this definition, a pattern in a text will be identified as a US state code if it corresponds to one of the strings between quotation marks, i.e. one of the definition items, such as AL or AK or WY (Note that the symbol "|" means "OR").

The structure #**7** gives a definition of what is to be identified as a street address. In this context, a street address is to be understood as a postal address excluding the name of the recipient. A typical example of a street address is: 225 Franklin Street, 02110 MA Boston. According to the definition given by structure #**7**, a pattern is a street address if it has elements matching the following sequence of definition items:

- a number in the sense as defined by structure #**4**, followed by
- some spaces, followed by
- some capitalized words, followed by,
- optionally, a known street type in the sense as defined by structure #**5** (the optional nature being indicated by the question mark behind the brackets surrounding "known_ street_type"), followed by
- a coma or spaces, followed by,
- optionally, a postal code in the sense as defined by structure #**3**, followed by
- some spaces, followed by
- a city in the sense as defined by structure #**6**.

This definition of a street address is deliberately broad in order to ensure that the application is able to identify not only street addresses written according to a single specific notation but also addresses written according to differing notations.

However, an application using such a broad definition is prone to the detection of a large number of false positives. For example, with the definition of a street address given above, the pattern "4 Apple Pies" will be wrongly recognized as a street address. The obvious solution to reduce the number of false positives is to make the structure definitions narrower. Yet, with narrow definitions there is an increased risk of missing interesting patterns.

At least certain embodiments of the present invention provide a method and system for identifying patterns in text using structures, which increase the flexibility of structure definitions and which, in particular, permit the formulation of structure definitions that lead to more accurate results during pattern identification.

SUMMARY OF THE DESCRIPTION

A computer-based method, in one embodiment, for identifying patterns in text using structures defining types of patterns which are to be identified, wherein a structure comprises one or more definition items, and wherein the methods include assigning a weighting to each structure and each definition item; searching the text for a pattern to be identified on the basis of a particular structure, a pattern being provisionally identified if it matches the definition given by said particular structure; in a provisionally identified pattern, determining those of the definition items making up said particular structure that have been identified in the provisionally identified pattern; combining the weightings of the determined definition items and optionally, the weighting of the particular structure, to a single quantity; assessing whether the single quantity fulfils a given condition; depending on the result of said assessment, rejecting or confirming the provisionally identified pattern.

Through the introduction of weightings for each structure and definition item, pattern definition and identification becomes more flexible and accurate. Indeed, in contrast to the conventional method of pattern identification, at least certain embodiments of a method of the invention introduce a supplementary test for the identification of patterns. It is no longer sufficient for a pattern to be recognized that it matches the definition of the corresponding structure. On top of that, at least certain embodiments of the invention use a second procedure which consists in performing a sort of plausibility check. The weightings of the definition items of the relevant structure that have been matched to the elements of the provisionally identified pattern must in combination fulfill a given condition. If this is the case, it is assumed that the identified pattern is sufficiently likely to really correspond to the relevant structure (e.g., if the structure defines telephone numbers, when the given condition is met by the combined weightings, it is assumed that the identified pattern is indeed a telephone number and not a false positive).

The introduction of weightings and of a probability test based on those weightings allows for structures with broad pattern definitions without the risk of an overly high number of false positives. A structure having a broad definition will lead to a lot of incorrect matches. However, these false positives may then be "sieved out" with the described "plausibility test" based on the assigned weightings. The weightings are assigned to the structures and definition items such that the combined weightings of a false positive are very unlikely to